Treatment of Missing Data in Hierarchically Structured Administrative Records:

A Case Study in the Bakken Region Using FBI's National Incident Based Reporting System

RTI International

Dan Liao

Marcus Berzofsky

David Heller

Kelle Barrick

Matthew DeMichele

Bureau of Justice Statistics

Kim Martin

Alexia Cooper

2015 FCSM Research Conference

December 2nd, 2015

Acknowledgements

The authors would like to thank BJS for sponsoring this research. However, we would like to note that the views expressed in this paper are those of the authors only and do not reflect the views or position of BJS or the Department of Justice.

Background

- Since 1929, the **Uniform Crime Reporting (UCR) Program** has collected information about crimes known to law enforcement and arrests on seven main offenses, and it started reporting on arson in 1979. Each month, the traditional **UCR Summary Reporting System (SRS)** collects counts of the number of these crimes known to law enforcement.

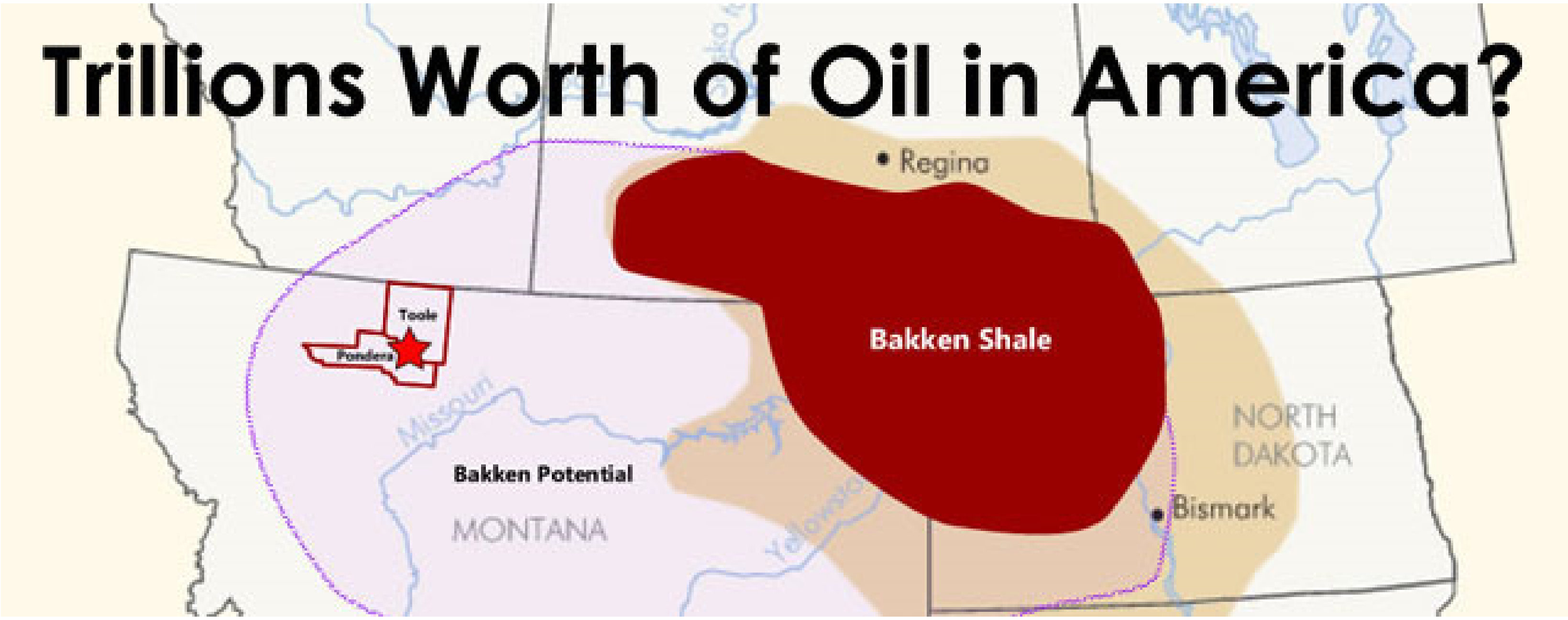
- With 1991 data, the UCR program began moving from summary counts to a more comprehensive and detailed reporting system known as the **National Incident-Based Reporting System (NIBRS).** For each crime incident coming to the attention of law enforcement, a variety of data are collected about the incident, including the nature and types of specific offenses in the incident, characteristics of the victim(s) and offender(s), the location of the incident, and characteristics of persons arrested in connection with a crime incident.

NIBRS Data

- As of May 2014, 32 states have been certified to report NIBRS to the FBI. 15 of them are 100% NIBRS reporters, meaning that all (or nearly all) of law enforcement agencies in the state submit only incident-based data to the NIBRS.

- Similar to other sources of administrative records, NIBRS data is plagued by missing data, which can cause significant bias in statistical estimation and obstructs analysts’ ability to make inferences directly from the data. In this paper, we propose an imputation method to deal with missing data in NIBRS.

Crime in the Bakken Region (2006-2012)



The purpose of this study is to use data from NIBRS and other available data sources to examine how crimes reported to the police, law enforcement responses to crime (arrests and clearances), and law enforcement staffing have changed in the Williston Basin/Bakken region as oil and natural gas production increased.

Multiple Data Sources

- **Main Data File: NIBRS**

  Montana, South Dakota and North Dakota are 100% reporters to NIBRS.

- **Auxiliary Data**
  - **UCR's SRS Data:** monthly crime counts by violent and property victimizations
  - **UCR's LEOKA** (Law Enforcement Officers Killed and Assaulted): characteristics of each agency (e.g. number of make/female officers)
  - **BJS's LEAIC** (Law Enforcement Agency Identifiers Crosswalk) file: agency type
  - **Annual Population Estimates from the US Census:** these estimates are available for all counties in the Bakken region in the period of interest and are disaggregated by sex and age groups

Missing Data in NIBRS

- NIBRS data is missing in a hierarchical structure
  - Item nonresponse at victim/incident level (e.g. weapon, injury, victim and offender relationship)
  - Unit nonresponse at agency level
    - ❑ some agencies do not report to NIBRS in the entire year
    - ❑ some agencies report to NIBRS in a partial year

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Agency1	√	√	√	√	√	√	√	√	√	√	√	√
Agency2	√			√					√	√	√	√
Agency3	√	√	√									
Agency4												

Unit Nonresponse: Weighting or Imputation

- Statistics of Interest
  - annual violent/property victimization rate (per 10K population) at county level
  - annual violent/property victimization rate by demographic groups at state level

- Calibration Weighting
  - unrealistic to do calibration at county level

    For example, if there is only one large agency in a county but it did not report to NIBRS, it could cause bias if we consider the small agencies are alike to this large agency in the same county and assign large weights to the small agencies to represent the large agency.
  - cause biased county-level estimates when not calibrating at county level

- Imputation: Hot Deck Imputation
  - find similar agencies in the data file and use them as donors

Hot Deck Imputation for Hierarchical Data

First Step: Imputation at Victim Level

item missingness in three variables: a) presence of weapon, b) injury sustained, and c) victim-offender relationship

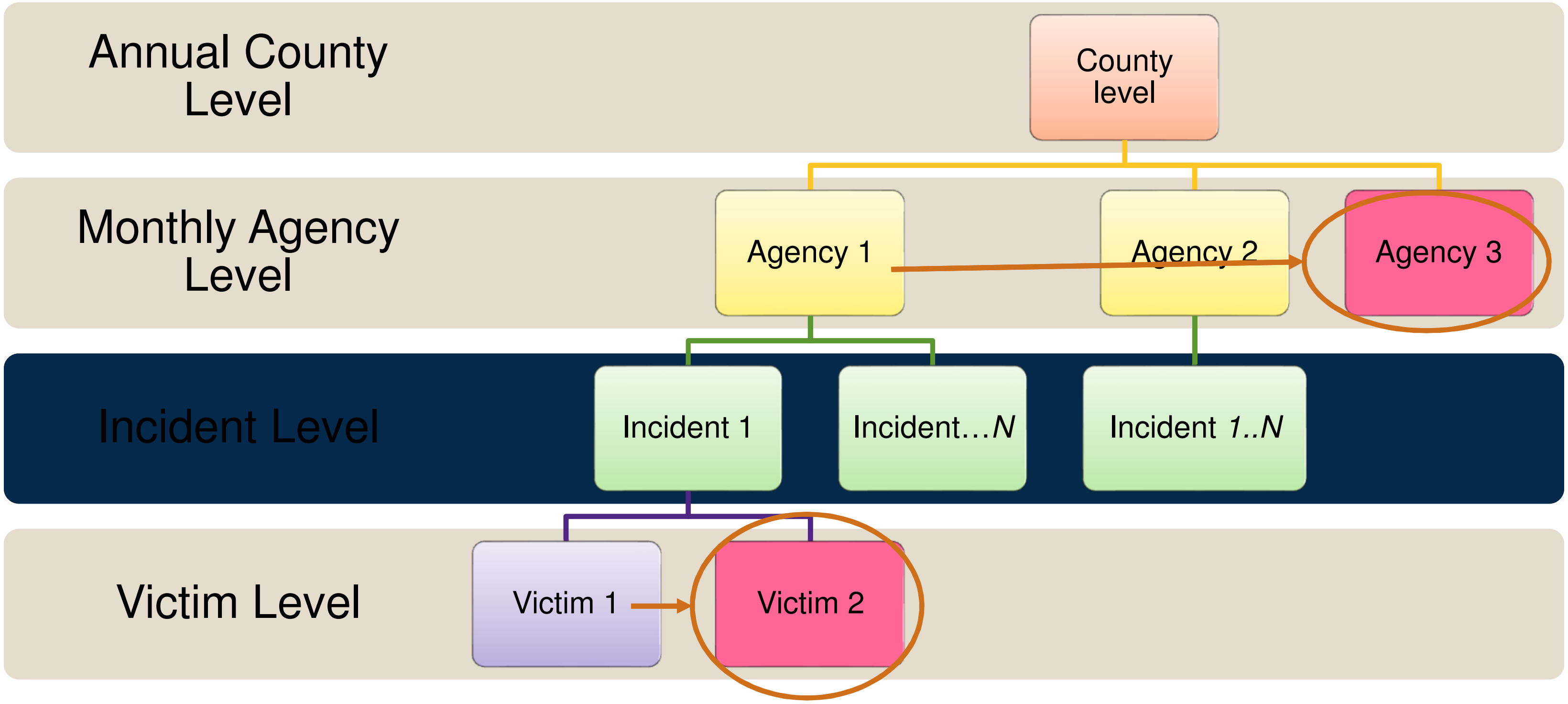
- hot deck imputation using matching variables to identify a donor for each missing value
- donor’s value is used to impute the missing value

Second Step: Imputation at Monthly-agency Level

Unit missingness due to nonresponse at the agency level

- hot deck imputation using matching variables to identify donors for each missing agency
- the entire record of the donor is used to impute the missing agency

Imputation Procedure for Data with a Hierarchical Structure



Annual County Level

County level

Monthly Agency Level

Agency 1

Agency 2

Agency 3

Incident Level

Incident 1

Incident…*N*

Incident *1..N*

Victim Level

Victim 1

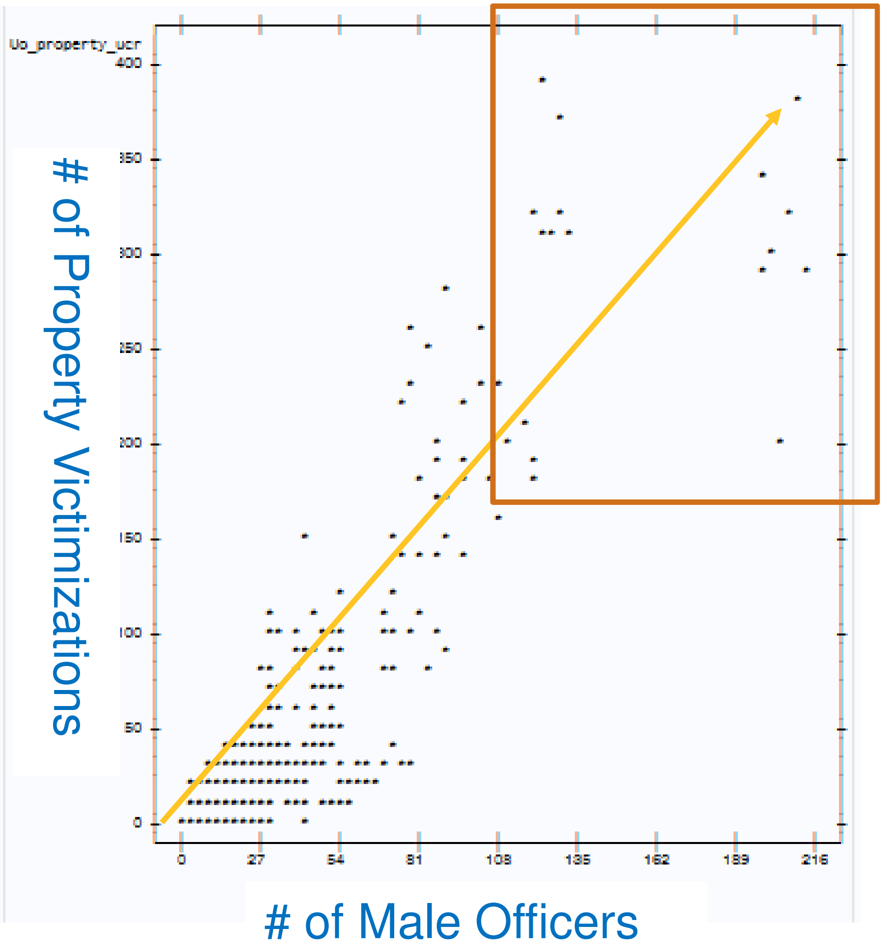
Victim 2

Matching Variables

- Data Merging: merge all the data sources in one big data file
- Potential Matching Variables
  - Victim Level: variables that are highly or moderately correlated with the variables to be imputed
    - ❑ Incident-specific variables
    - ❑ agency-specific variables
  - Agency Level: variables that are highly or moderately correlated with the variables of interest (e.g. counts of violent/property victimizations)
    - ❑ agency-specific variables

- Selection of Final Matching Variables
  - Categorical variable: two-way frequency tables
  - Continuous variable: scatterplots
    - Scatterplots are used rather than regression models in case of outliers or nonlinear relationship.

An Example of Scatterplot



Matching Variables

- Selected matching variables
  - Victim Level

    Oil Boom Period Onset; State; type of victim (individual/police officer); victim’s age; victim’s race; offender’s age; offender’s race; victim’s gender; offender’s gender; victim’s ethnicity

  - Agency Level

    Boom Period Onset; State (MT, ND, SD); indicator of metropolitan statistical area; agency type; different population size groups; and agency groups with more or less male officers

- An algorithm was developed to merge small donor groups

Use of Auxiliary Data

- LEOKA and LEAIC data were used to create a complete list of agencies in MT, ND and SD from 2006 through 2012. If an agency on this list is not listed in the NIBRS data at the agency level, this agency will be considered missing.

- UCR' SRS, LEOKA and LEAIC data were considered as potential matching variables in imputation.

- UCR's SRS data were used for selecting final matching variables.
  - Number of Property Victimizations vs. Number of Male Officers

- UCR's SRS data: used to identify zero-crime agency at monthly level

- Annual population estimates from the US Census were used in conjunction with the NIBRS data to calculate the victimization rate per 10k population.

Multiple Imputation

- For each imputed variable, we impute it 5 times at victim level (*j*) and 5 times at monthly-agency level (*i*), which makes 25 imputed datasets at total. The mean of the estimates (e.g. total crime counts) derived from the 25 imputed datasets is used as the final estimate

$$\bar{\theta} = \frac{1}{25}\sum_{i=1}^{5}\sum_{j=1}^{5}\hat{\theta}_{i(j)}$$

- **Variance Estimation**

Total Variance Estimator (Var$_{25,}$ CV$_{25}$) for 25 imputed dataset

$$T = \bar{U} + \left(1 + \frac{1}{25}\right)B$$

where $\bar{U}$ is the average of the 25 imputed variances (“**within imputation**” component) and $B = (25-1)^{-1}\sum_{i=1}^{5}\sum_{j=1}^{5}\left(\hat{\theta}_{i(j)} - \bar{\theta}\right)^{2}$ (**“across imputation”** component).

From design-based perspective, **the imputed variance within each imputed dataset is equal to 0,** because the dataset we are dealing with (NIBRS data) is from administrative records (not a probability sample of the finite population). Therefore, $\bar{U} = 0$.

Presence of Weapons in Violent Crime Victimizations

Region	Year	Rate (%)	CV$_{25}$
Bakken	**2006**	48.59	0.73%
	2007	51.74	1.55%
	2008	49.51	0.42%
	2009	48.17	0.45%
	2010	42.42	0.42%
	2011	47.22	0.74%
	2012	46.72	0.34%
Non-Bakken	**2006**	46.40	0.18%
	2007	46.63	0.27%
	2008	47.79	0.23%
	2009	49.17	0.24%
	2010	48.81	0.23%
	2011	48.98	0.20%
	2012	48.98	0.11%

Violent Victimization Rates (per 10k population), by Victim and Offender Relationship

Victim and Offender Relationship	Year	Rate (%)	CV_{25}
Stranger	**2006**	5.57	4.01%
	2012	6.93	1.28%
Intimate Partner	**2006**	24.34	0.72%
	2012	27.60	0.43%
Acquaintance	**2006**	20.98	1.11%
	2012	19.57	0.61%
Family	**2006**	10.65	2.42%
	2012	11.63	0.97%
Other	**2006**	4.07	3.27%
	2012	5.45	0.88%
Unknown	**2006**	0.00	--
	2012	0.97	2.98%

Challenges and Future Research

- Data Volume when Dealing with Administration Records

- Nonresponse Assessment
  - Unit nonresponse in NIBRS
    - Why some agencies do not report to NIBRS at all?
    - Why some agencies do not report to NIBRS regularly?
  - No crime in a month (zero crime agencies)

- Data Editing
  - Missing data can be caused by the setup of the system. For example, injury sustained variable is missing for all the murder incidence.

Challenges and Future Research (con't.)

- Experimenting Different Imputation Methods
  - Imputing at agency level first and then at victim level
  - Regression tree and other statistical methods to select matching variables
- Expanding the Usage of NIBRS Data

More Information

Dan Liao

Research Statistician

301.816.4615

dliao@rti.org